Reinforcement Learning	March 13, 2018
Recitation 3	
Lecturer:Yishay Mansour	*TA: Lee Cohen*

Part 1: Markov Chain

Definition 1. Markov Chain is a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$

- $\mathcal{S}$ is a (finite) set of states.
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s]$.

Note that each row of $\mathcal{P}_{ss'}$ sums to 1.

Definition 2. Let $p_{ij}^{(m)} := \mathcal{P}(X_m = j | X_0 = i)$.

- State j is **accessible** from i (denoted by $i \to j$) if $p_{ij}^{(m)} > 0$ for some $m \geq 1$.
- State $i \in \mathcal{S}$ has **period** of $d_i := GCD\{m \geq 1 : p_{ii}^{(m)} > 0\}$.
- State i is a-periodic if $d_i = 1$.
- State $i \in \mathcal{S}$ is **recurrent** if $\mathcal{P}(X_t = i \text{ for some } t \geq 1 | X_0 = i) = 1$. Otherwise, state i is transient.

Definition 3. States i and j are **communicating** if: $(i \to j) \wedge (j \to i)$.
I.e. j is accessible from i and i is accessible j.

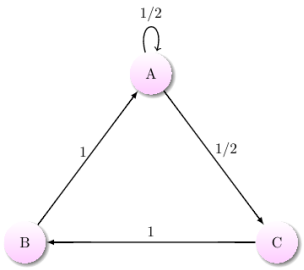
Definition 4. Markov chain $\langle \mathcal{S}, \mathcal{P} \rangle$ is **irreducible** if all states communicate with each other. I.e., the graph that is induced from $\langle \mathcal{S}, \mathcal{P} \rangle$ is strongly connected.

Definition 5. Let T_i be the return time to state i (i.e., T_i is the number of stages required for (X_t) to return to i).

1. If i is a recurrent state, then $T_i < \infty$ w.p. 1.
2. State i is **positive recurrent** if $E(T_i) < \infty$.
3. State i is **null recurrent** if $E(T_i) = \infty$.
4. If the state space is finite, all recurrent states are positive recurrent.

Definition 6. The probability vector $\pi = (\pi_i)$ is a **stationary/steady state** distribution for the Markov Chain if $\pi \mathcal{P} = \pi$.

Example 1. In the following Markov Chain, all states are aperiodic and positive recurrent.



From the Theorem shown in class - irreducible aperiodic Markov Chain have a *unique stationary distribution* π^*. It follows:

$$P = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \pi P = \pi \tag{1}$$

To obtain:

$$\pi^* = (0.5, 0.25, 0.25) \tag{2}$$

Exercise 1. Let $\{X_n\}$ be a time-homogenous Markov processes in discrete time which takes values in $\{0, 1, ...\}$ (an infinite countable set).

1. Assume the process satisfies for each $i \in \{0, 1, ...\}$:

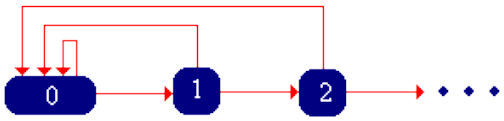
$$p_{i,0} = q, \quad p_{i,i+1} = 1 - q, \quad 0 < q < 1. \tag{3}$$

   Plot the state transition diagram for {0,1,2,3}. If the chain is recurrent, find the stationary distribution, if it is not find the transient states.

2. Consider the same process as above and assume $P(X_0 = 0) = 1$. Define $Y_n = |\{\tau : X_\tau = 0, \tau \leq n\}|$ as the number of visits to 0 until time n. Also define $Z_n = (X_n \quad Y_n)^T$. Is $\{Z_n\}$ a Markov process? Is it recurrent? Is it transient?

Solution 1.

1.



First we note that:

(a) The chain is irreducible, since 0 leads to every other state and every state leads back to 0.

(b) The chain is aperiodic since state 0 period is 1, and Periodicity is a class property.

Next, we see that the chain is recurrent if and only if a failure is sure to occur (return to state 0).

$$P[\exists t \geq 1 : X_t = 0 | X_0 = 0] = q\sum_{i=0}^{\infty}(1-q)^i = \frac{q}{1-(1-q)} = 1 \tag{4}$$

Therefore, state 0 is positive recurrent and since the chain is irreducible and aperiodic - all the states are positive recurrent. Finally, we find the stationary distribution:

$$P = \begin{bmatrix} q & 1-q & 0 & 0 & \dots \\ q & 0 & 1-q & 0 & \dots \\ q & 0 & 0 & 1-q & \dots \\ \vdots & & \ddots & & \end{bmatrix}, \quad \pi P = \pi \tag{5}$$

$$\pi_0 = \sum_{i=0}^{\infty}\pi_i q, \quad \pi_i = \pi_{i-1}(q-1) \quad \forall i \in \{1,2,3,...\} \tag{6}$$

$$\pi_0 = q\sum_{i=0}^{\infty}\pi_i = q, \quad \pi_i = q(q-1)^i \quad \forall i \in \{1,2,3,...\} \tag{7}$$

2. First we show Z_n is indeed a Markov Chain.

$$P(Z_{t+1} = \binom{i+1}{j} | Z_t = \binom{i}{j}) = 1-q \tag{8}$$

$$P(Z_{t+1} = \binom{0}{j+1} | Z_t = \binom{i}{j}) = q \tag{9}$$

$$P(Z_{t+1} = \binom{1}{j} | Z_t = \binom{0}{j}) = 1 \tag{10}$$

Next, we observe that all the states in the chain are transient, since no state leads to it's predecessor state (the chain has multiple connectivity components and is aperiodic).

Part 2: Markov Decision Process

Definition 7. Markov Decision Process is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states.
- $\mathcal{A}$ is a finite set of actions.
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$.
- $\mathcal{R}$ is a reward function, $\mathcal{R}^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$.
- γ is a discount factor $\gamma \in [0, 1]$.

It is a Markov Chain with rewards and actions.

Definition 8. A **policy** π is distribution over actions given states

$$\pi(a|s) = P[A_t = a | S_t = s]. \tag{11}$$

Definition 9. The **state-value function** $V^\pi_k(s)$ of an MDP is the expected return starting from state s, and then following policy π

$$V^\pi_k(s) = E_\pi[\sum_{t=k}^{T} R_t | S_k = s] \tag{12}$$

Exercise 2. In **Micro-Blackjack**, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your utility is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you Draw, you receive no utility. There is no discount ($\gamma = 1$).

1. What are the states and the actions for this MDP?
2. What is the transition function and the reward function for this MDP?
3. Give the optimal policy for this MDP.

Solution 2.

1. The state is the current sum of your cards, plus a terminal state. There are 2 actions - draw and stop:

$$\mathcal{S} = \{0, 2, 3, 4, 5, \text{Done}\} \quad \mathcal{A} = \{\text{Draw}, \text{Stop}\} \tag{13}$$

2. The transition function is

$$\mathcal{P}(s, \text{Stop}, \text{Done}) = 1 \tag{14}$$

$$\mathcal{P}(s, \text{Draw}, s') = \begin{cases} 1/3 & \text{if } s' - s \in \{2, 3, 4\} \\ 1/3 & \text{if } s = 2 \text{ and } s' = \text{Done} \\ 2/3 & \text{if } s = 3 \text{ and } s' = \text{Done} \\ 1 & \text{if } s \in \{4, 5\} \text{ and } s' = \text{Done} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

The reward function is

$$\mathcal{R}(s, a, s') = \begin{cases} s & \text{if } s \leq 5 \text{ and } a = \text{Stop} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

3. In general, for finding the optimal policy for an MDP, we would use some method like value iteration followed by policy extraction. However, in this particular case, it is simple to work out that the optimal policy would be Draw if $s \leq 2$, Stop otherwise.

Part 3: Finite Horizon MDP

Theorem 1. Finite-horizon Dynamic Programming. The following holds:

1. Backward recursion: Set $V_T(s) = r_T(s)$ for $s \in S_T$.
   For $k = T-1, ..., 0$, $V_k(s)$ may be computed using the following recursion:

$$V_k(s) = \max_{a \in A_k} \{r_k(s,a) + \sum_{s' \in S_{k+1}} p_k(s'|s,a) V_{k+1}(s')\}, s \in S_k \tag{17}$$

2. Optimal policy: Any Markov policy that satisfies, for $t = 0, ..., T-1$,

$$\pi_t^*(s) \in \arg\max_{a \in A_t} \{r_t(s,a) + \sum_{s' \in S_{t+1}} p_t(s'|s,a) V_{t+1}(s')\}, s \in S_t \tag{18}$$

   is an optimal control policy. Furthermore, maximizes $V_0(s_0)$ simultaneously for every initial state $s_0 \in S_0$.

Note that the theorem species an optimal control policy which is a deterministic Markov policy.

Exercise 3. Suppose an investor has 10,000$ to his disposal and he must decide how to invest it, so as to maximize his total expected returns. The investor may choose between investing all of his capital either in stock from company A or in stock from company B. Investing in company A renders a profit of 100% (i.e. his investment is doubled) after one year with probability 0.10. With probability 0.90 however, there will no profit after one year, and the investor will get his investment back. Company B has a higher risk profile, but renders higher expected returns. With probability 0.6 the investment is doubled, whereas with probability 0.4 the investment is completely lost. As a consequence, the expected profit from investing 10,000$ in company A is

$$0.10 \cdot 10,000 = 1,000 \tag{19}$$

and in company B this is

$$0.60 \cdot 10,000 + 0.4 \cdot (-10,000) = 2,000. \tag{20}$$

If not bankrupt, the investor can re-invest his money every year (each time 10,000 due to the popularity of both investments). What is the best investment strategy for the investor, if his goal is to maximize the expected profit after 5 years?

Solution 3. First, we will define the system as MDP:

$$\begin{gathered}
\mathcal{S} = \{0, 10000, 20000, 30000, 40000, 50000, 60000\} \\
\mathcal{A}(s) = \{A, B\}, s \neq 0 \\
\mathcal{A}(0) = \{0\} \\
\mathcal{P}(s'|s = n, a = A) = \begin{cases} 0.1 & s' = n + 10,000 \\ 0.9 & s' = s \end{cases} \\
\mathcal{P}(s'|s = n, a = B) = \begin{cases} 0.6 & s' = n + 10,000 \\ 0.4 & s' = s - 10,000 \end{cases}
\end{gathered} \tag{21}$$

Note that the rewards depends on the next stage, therefore we will use the **expected reward**:

$$\begin{gathered}
\mathcal{R}(s, A) = 0.1 \cdot 10,000 + 0.9 \cdot 0 = 1,000 \\
\mathcal{R}(s, B) = 0.6 \cdot 10,000 + 0.4 \cdot -10,000 = 2,000 \\
\mathcal{R}(0, *) = 0 \\
\mathcal{R}_T(s) = 0
\end{gathered} \tag{22}$$

Now we can apply the theorem above to retrieve the optimal investment policy. We will begin with $\underline{t = 4}$, and go backward:

$$\begin{gathered}
V_4(s) = \max\{1000 + 0.1 \cdot 0 + 0.9 \cdot 0, 2000 + 0.6 \cdot 0 + 0.4 \cdot 0\} = 2000, \\
\pi_4^*(s) = B \quad s \in \{10000, ..., 50000\}
\end{gathered} \tag{23}$$

$$V_4(0) = 0$$

$\underline{t=3}$:

$$\begin{gathered} V_3(s) = \max\{1000 + 0.1 \cdot 2000 + 0.9 \cdot 2000, 2000 + 0.6 \cdot 2000 + 0.4 \cdot 2000\} = 4000, \\ \pi_3^*(s) = B \quad s \in \{20000, ..., 40000\} \\ V_3(10000) = \max\{1000 + 0.1 \cdot 2000 + 0.9 \cdot 2000, 2000 + 0.6 \cdot 2000 + 0.4 \cdot 0\} = 3200, \\ \pi_3^*(10000) = B \end{gathered} \tag{24}$$

$\underline{t=2}$:

$$\begin{gathered} V_2(30000) = \max\{1000 + 0.1 \cdot 4000 + 0.9 \cdot 4000, 2000 + 0.6 \cdot 4000 + 0.4 \cdot 4000\} = 6000, \\ \pi_2^*(30000) = B \\ V_2(20000) = \max\{1000 + 0.1 \cdot 4000 + 0.9 \cdot 4000, 2000 + 0.6 \cdot 4000 + 0.4 \cdot 3200\} = 5680, \\ \pi_2^*(20000) = B \\ V_2(10000) = \max\{1000 + 0.1 \cdot 4000 + 0.9 \cdot 3200, 2000 + 0.6 \cdot 4000 + 0.4 \cdot 0\} = 4400, \\ \pi_2^*(10000) = B \end{gathered} \tag{25}$$

$\underline{t=1}$:

$$\begin{gathered} V_1(20000) = \max\{1000 + 0.1 \cdot 6000 + 0.9 \cdot 5680, 2000 + 0.6 \cdot 6000 + 0.4 \cdot 4400\} = 7360, \\ \pi_1^*(20000) = B \\ V_1(10000) = \max\{1000 + 0.1 \cdot 5680 + 0.9 \cdot 4400, 2000 + 0.6 \cdot 5680 + 0.4 \cdot 0\} = 5528, \\ \pi_1^*(10000) = A \end{gathered} \tag{26}$$

$\underline{t=0}$:

$$\begin{gathered} V_0(10000) = \max\{1000 + 0.1 \cdot 7360 + 0.9 \cdot 5528, 2000 + 0.6 \cdot 7360 + 0.4 \cdot 0\} = 6711.20, \\ \pi_0^*(10000) = A \end{gathered} \tag{27}$$